SEQRITE

AI/ML in Cybersecurity

WHITE PAPER

09 Sep, 2025

www.seqrite.com

Table of Contents

1. Executive Summary

Artificial Intelligence (AI) and Machine Learning (ML) are catalysing a paradigm shift in cybersecurity, fundamentally altering the landscape for both defensive and offensive operations. This white paper provides a comprehensive analysis of this transformation, charting the historical evolution of AI in security, detailing its contemporary applications, and examining the profound implications for the future of digital defence.

The white paper traces the journey from early, deterministic rule-based systems of the pre-2000 era to the data-driven machine learning models of the 2000s, which revolutionised fields like spam filtering and intrusion detection. The subsequent decade saw the rise of deep learning, enabling advanced malware analysis and the integration of AI into Security Orchestration, Automation, and Response (SOAR) platforms. The current era, dominated by Generative AI (GenAI), has further accelerated this evolution, equipping defenders with tools for rapid threat intelligence analysis and automated response, while simultaneously arming adversaries with the ability to craft sophisticated and highly personalised attacks at an unprecedented scale.

A central theme of this white paper is the dual-use nature of AI. While AI/ML offers significant benefits, including enhanced threat detection, predictive analytics, and operational efficiency, it also introduces new and formidable challenges. Attackers now leverage AI to automate reconnaissance, generate polymorphic malware, and create convincing deepfakes for social engineering campaigns. This has lowered the barrier to entry for less-skilled actors and shifted the traditional asymmetry of cyber conflict, turning it into a battle of competing AI capabilities.

Furthermore, AI systems themselves have become a new attack surface. This white paper examines the vulnerabilities inherent in the ML pipeline, as categorised by the MITRE ATLAS framework, including data poisoning, model evasion, and sensitive data leakage from training sets. These challenges underscore a critical reality: securing an organisation in the age of AI requires not only leveraging **AI for security** but also mastering the principles of the **security of AI**.

To navigate this complex environment, the white paper details essential governance and implementation frameworks. Gartner's AI Trust, Risk, and Security Management (AI TRiSM) and the National Institute of Standards and Technology's (NIST) AI Risk Management Framework (RMF) provide structured approaches for embedding trust, safety, and security throughout the AI lifecycle. Implementing these guardrails, guided by maturity models and measured by clear performance metrics like Mean Time to Detect (MTTD) and Mean Time to Respond (MTTR), is crucial for responsible AI adoption.

India's cybersecurity posture is further strengthened by indigenously developed platforms like Seqrite, which embed advanced AI/ML-driven threat detection and autonomous response, ensuring rapid, localised protection for enterprises.

In conclusion, AI is not a silver bullet but a foundational technology that demands a strategic, lifecycle-based approach to security. The path forward requires a holistic strategy that embraces AI's defensive potential while proactively mitigating its inherent risks. For organisations to remain resilient, they must foster a culture of continuous adaptation, where human expertise is augmented – not replaced – by secure, transparent, and trustworthy AI systems.

2. History of AI/ML in Cybersecurity

The history of Artificial Intelligence in cybersecurity is not a simple timeline of technological advancement but rather an escalating arms race. Each evolution in defensive AI has been met with corresponding innovations in offensive tactics, creating a feedback loop that has dramatically accelerated the pace of change in the digital threat landscape. This chapter traces that evolution through four distinct eras.

2.1. Pre - 2000

This foundational period was dominated by symbolic AI, where intelligence was not learned from data but explicitly encoded as rules and logical structures by human experts. The primary goal was to codify and automate human expertise to perform specific, well-defined security tasks.

2.1.1. Good Cases

- **Expert Systems for Intrusion Detection**: The most significant development of this era was the Intrusion Detection Expert System (IDES), developed at SRI International during the 1980s. IDES was a pioneering expert system that monitored audit logs from mainframe computers in real-time. It used a combination of user-defined rules to detect known misuse patterns and statistical models to identify anomalous behaviour that deviated from established user profiles. This represented a critical first step toward automated, continuous security monitoring. [1]
- **Rule-Based Systems**: Early cybersecurity tools such as firewalls, antivirus software, and access control systems were fundamentally rule-based systems. These tools operated on a set of predefined "if-then" conditions. Firewalls, for instance, used rules based on IP addresses, ports, and protocols to permit or deny network traffic. Antivirus programs relied on a database of "signatures", unique patterns found in known malware, to identify and block malicious files. These systems were effective against known threats and formed the bedrock of enterprise security for decades.

2.1.2. Bad Cases

- **Rigidity and Inability to Detect Novel Threats**: The primary weakness of rule-based systems was their static nature. They were completely unable to detect new or "zero day" attacks for which no predefined rule or signature existed. As the threat landscape grew, the reliance on manual updates from human experts became a significant bottleneck, making it impossible to keep pace with the emergence of new malware.
- **Evasion through Polymorphism**: Attackers quickly learned to exploit the rigidity of signature-based detection. In the early 1990s, polymorphic viruses appeared. This type of malware could alter its own code with each new infection, creating a unique signature every time while keeping its malicious payload intact. This simple technique rendered signature-based antivirus software largely ineffective, forcing the security industry to look beyond static rules.

2.2. 2000 - 2010

Fueled by Moore's Law and the explosion of data from the internet, this decade saw a pivotal shift from symbolic AI to a data-driven paradigm. Machine learning (ML)models, which learn patterns directly from data, began to offer a more dynamic and adaptive approach to cybersecurity.

2.2.1. Summary

The period was defined by the practical application of statistical machine learning to solve large-scale security problems that were intractable for rule-based systems. Spam filtering and network anomaly detection became the flagship use cases, demonstrating the power of learning from data to combat evolving threats.

2.2.2. Good Cases

- **Machine Learning for Spam Filtering**: The fight against unsolicited bulk email (spam) became one of the first major success stories for ML in cybersecurity. Algorithms like Naive Bayes, which calculates the probability that an email is spam based on the words it contains, and Support Vector Machines (SVMs), were deployed with great success. These models could learn the statistical properties of spam and adapt to new campaigns far more effectively than static keyword filters. Google's implementation of ML in Gmail's spam filter was a landmark achievement, demonstrating that ML could achieve extremely high accuracy and a low false-positive rate at a massive scale.

- **Machine Learning in Intrusion Detection**: Researchers began applying neural networks and other ML models to Intrusion Detection Systems (IDS). These systems were trained on network traffic data to learn a baseline of "normal" activity. They could then identify anomalies – deviations from this baseline – that might indicate a previously unseen attack. This marked a significant step beyond signature-based detection toward identifying novel threats. [2]

2.2.3. Bad Cases

- **High False Positive Rates**: A persistent problem for early anomaly detection systems was the high rate of false positives. Legitimate but unusual network behaviour, such as an administrator running a rare diagnostic tool, could be flagged as malicious. This created "alert fatigue" among security analysts, who were inundated with alerts that were not actual threats, causing them to potentially miss genuine incidents.
- **Early Adversarial Evasion**: Spammers began to evolve their tactics specifically to deceive ML filters. Techniques included "word salading" (inserting random legitimate words into spam emails) and misspelling malicious keywords to evade statistical detection. This marked the beginning of an adversarial dynamic where attackers targeted the logic of the ML models themselves.
- **Significant AI Failures**: The stakes of AI errors in security were starkly illustrated by a 1983 incident where a Soviet nuclear early-warning system falsely reported an incoming missile strike from the United States. The decision of a human officer, Stanislav Petrov, to defy protocol and report it as a system malfunction prevented a potential nuclear war, highlighting the catastrophic potential of false positives in automated security systems. [3]

2.3. 2010 - 2020

This decade was defined by the deep learning revolution and the operationalisation of AI within mainstream security platforms. AI transitioned from a specialised tool to a core component of the modern Security Operations Centre (SOC). However, this deep integration also revealed a new and dangerous attack surface: the AI models themselves

2.3.1. Summary

Deep learning models, with their ability to learn from vast, unstructured datasets, provided breakthroughs in areas like malware analysis. AI became central to a new category of tools – SOAR platforms – that automated security workflows. Concurrently, the field of adversarial machine learning matured, as attackers began to systematically exploit the inherent vulnerabilities of ML models.

2.3.2. Good Cases

- **Deep Learning for Anomaly Detection**: Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), were successfully applied to anomaly detection. One innovative technique involved visualising a malware binary's code as a grayscale image. CNNs could then identify textural and structural patterns in these images that were characteristic of specific malware families. This approach proved highly effective against obfuscated and novel malware, with academic studies reporting detection accuracies of over 95%. [4]
- **AI in SOAR Platforms**: The mid-2010s saw the emergence of Security Orchestration, Automation, and Response (SOAR) platforms. These tools began incorporating AI analytics to automatically enrich alerts with threat intelligence, prioritise incidents based on risk, and orchestrate response actions. This significantly reduced manual effort for SOC analysts and decreased response times.

2.3.3. Bad Cases

- **Adversarial Machine Learning Attacks**: As ML models became central to security, attackers shifted their focus to exploiting them directly. This gave rise to the formal discipline of adversarial machine learning. Two primary attack types were identified:
- **Evasion Attacks**: An attacker makes small, often imperceptible, perturbations to a malicious input to cause it to be misclassified by a model at inference time. For example, slightly altering a few pixels in a malware binary's image representation could cause a CNN-based detector to classify it as benign.
- **Poisoning Attacks**: An attacker with access to the training data injects malicious samples to corrupt the learning process. This can be used to create a "backdoor" in the model, causing it to misclassify specific inputs, or simply to degrade its overall performance.

2.4. 2020 - Present

The current era is being shaped by the widespread availability of powerful Generative AI, especially Large Language Models (LLMs). This has initiated another profound shift, where AI is no longer j ust an analytical or predictive tool but a generative one. This capability is being leveraged for both highly advanced cyber defence and offense, collapsing the cycle time between innovation and exploitation.

2.4.1. Summary

Generative AI is being integrated into security platforms to act as an intelligent assistant for analysts, automating complex tasks like threat hunting and reporting. At the same time, adversaries are using the same technology to generate highly sophisticated social engineering attacks, create polymorphic malware, and lower the barrier to entry for complex cybercrime, creating a new and dynamic threat landscape.

2.4.2. Good Cases

- **Generative AI for Threat Intelligence and Hunting**: LLMs can now process and synthesise vast quantities of unstructured data from threat intelligence reports, security blogs, and dark web forums. This allows them to automatically extract Indicators of Compromise (IOCs), identify attacker Tactics, Techniques, and Procedures (TTPs), understand the Indicators of Attack (IoA), Indicators of Behavior (IoB) and even generate threat hunting queries for security analysts, dramatically reducing research time.
- **AI-Powered Autonomous Response**: Modern security platforms now use self learning AI to provide real-time, autonomous response. These systems can detect and neutralise novel threats within seconds, often without any human intervention. For instance, upon detecting a ransomware attack in progress, the system can automatically sever the malicious network connections of the compromised device to contain the threat instantly.
- **Reduced Data Breach Costs**: The impact of AI in security is quantifiable. According to a 2025 IBM report, organisations that extensively use security AI and automation experience data breach costs that are, on average, USD 1.9 million lower than organisations that do not. This is attributed to significantly faster breach identification and containment times. [5]

2.4.3. Bad Cases

- **AI-Powered Offensive Campaigns**: Adversaries are now using GenAI as a force multiplier. AI Bots are leveraged for orchestrating large scale botnet and DDoS attacks. LLMs are used to generate highly convincing and personalised phishing emails at scale, overcoming the language and grammatical errors that often betrayed older campaigns. Deepfake technology is used to create voice clones of executives to authorise fraudulent wire transfers, a tactic known as Business Email Compromise (BEC).
- **Democratisation of Cybercrime**: Generative AI tools have significantly lowered the technical barrier to entry for cybercrime. Novice actors can now use LLMs to generate malicious code, create phishing websites, and plan complex operations that would have previously required years of expertise. Malicious LLMs like WormGPT and FraudGPT, which are specifically designed without ethical safeguards, are now available on dark web forums to facilitate these activities.

The history of AI in cybersecurity reveals an accelerating cycle of innovation and exploitation. Initially, the gap between the introduction of a new defensive technology, like signature-based antivirus, and its widespread evasion by attackers, such as through polymorphic malware, spanned several years. The subsequent shift to ML based anomaly detection was met more quickly with adversarial attacks targeting the models themselves. With the advent of Generative AI, this cycle has collapsed. The same foundational models that power defensive tools are simultaneously being used to enhance offensive capabilities. This near-instantaneous feedback loop between defence and offense signifies a fundamental change, demanding that security strategies evolve from periodic adaptation to a state of continuous, real-time resilience.

3. AI/ML Use Cases in Cybersecurity

Beyond its historical evolution, the practical application of traditional Artificial Intelligence (AI) and Machine Learning (ML) has become foundational to modern cybersecurity defence. These technologies have moved beyond academic research to become integral components of security stacks, delivering tangible benefits in threat detection, operational efficiency, and risk management. This chapter explores the primary benefits, use cases, and real-world applications of AI/ML in defending enterprise environments.

3.1. Benefits

The integration of AI/ML into cybersecurity operations offers three core advantages over traditional, manual, or purely rule-based approaches.

- **Enhanced Threat Detection and Analysis**: The foremost benefit of AI/ML is its ability to analyse vast and complex datasets at a scale and speed unattainable by human analysts. ML algorithms can sift through terabytes of network traffic, system logs, and endpoint data to identify subtle patterns, correlations, and anomalies that are often indicators of compromise. This capability allows for the detection of threats earlier in the attack lifecycle, including novel and previously unseen attacks that do not have known signatures
- **Predictive Analytics for Proactive Defence**: By training on historical data of past attacks and vulnerabilities, ML models can develop predictive capabilities. They can forecast potential future attack vectors, identify systems that are most likely to be targeted, and predict emerging threats before they are widely deployed. This shifts the security paradigm from a reactive posture, where teams respond to attacks after they occur, to a proactive one, where defences are strengthened against anticipated threats
- **Automation and Operational Efficiency**: AI automates many of the repetitive and time-consuming tasks that burden Security Operations Centres (SOCs), such as initial alert triage, data enrichment, and basic incident response actions. By filtering out the noise of false positives and prioritising genuine threats, AI frees up human analysts to concentrate on more complex, strategic investigations and threat hunting. This not only improves efficiency but also helps mitigate the effects of the global cybersecurity skills shortage

3.2. Use Cases

AI/ML is applied across a wide spectrum of security domains to address concrete detection, prioritization, and response challenges at scale.

- **Network Security and Intrusion Detection**: ML-powered NIDS establish baselines of normal behavior and surface deviations indicative of C2, exfiltration, and DDoS in near real time.
- **Malware Detection**: Deep models perform static and dynamic analysis to identify malicious traits and behaviors, enabling detection of zero-day variants beyond signatures.
- **Fraud Detection**: Real-time models score transactions across rich behavioral features to flag card fraud, scams, and illicit activity more accurately than rules.
- **Vulnerability Management**: AI prioritizes remediation by combining severity with asset criticality and exploit likelihood, shifting from reactive patching to risk-based action.

- **Security Orchestration, Automation, and Response (SOAR)**: Automation, and Response (SOAR): AI enriches alerts, ranks risk, and triggers playbooks (e.g., isolate endpoints), reducing MTTD and MTTR materially.
- **AI-Powered Endpoint Detection & Response**:

3.2.1. Seqrite Use Cases

Seqrite XDR applies analytics and machine learning across a unified, multi-layered security stack to detect anomalies, hunt threats, classify attacks with MITRE ATT&CK, and reduce false positives while providing playbook automation and a single investigative view.

It supports near real-time IOC-based hunting, consolidates related alerts into incidents, and exposes operational insights to streamline SOC workflows at scale.

- Real-time anomaly detection using machine learning algorithms to flag unusual patterns and enable swift remediation across endpoints, networks, cloud, and identities.
- IOC and TTP-driven threat hunting with saved searches, frequency analysis, and incident correlation to accelerate investigations.
- Automated alert prioritization and response to shrink dwell and response times, delivering 40–70% fewer false positives than traditional SIEM analytics.
- Playbook-based manual and automatic response orchestration to streamline triage, containment, and remediation.
- Multi-layered protection against zero-day threats with unified telemetry, effective attack classification, and MITRE ATT&CK mapping.
- Deep learning, behavioral analytics, and predictive intelligence (GoDeep.AI) that shorten breach response cycles by 108 days versus conventional methods (Seqrite internal benchmark).

6], [7]

3.3. Case Studies

The value of AI/ML in cybersecurity is best illustrated through its application in real-world scenarios by both specialised security vendors and organisations in other industries.

- **Darktrace**: Darktrace is a cybersecurity company whose platform is built around a core of "Self-Learning AI." Instead of relying on predefined rules or signatures, its AI engine learns the unique "pattern of life" for every user, device, and system in an organisation's digital environment. It then uses anomaly detection to identify deviations from this learned baseline.
- **Solution in Action**: In a case with pharmaceutical manufacturer CordenPharma, Darktrace's AI detected and autonomously blocked a crypto-mining attack during its trial period. For wealth services firm Aviso, the platform reduced a flood of alerts to just 73 actionable incidents by investigating 23 million events, allowing the security team to focus on genuine threats. This approach allows it to detect novel, never before-seen threats and respond autonomously to contain them in seconds

[8]

4. GenAI Use Cases in Cybersecurity

While traditional AI/ML excels at analysis and prediction, the emergence of Generative AI (GenAI), particularly Large Language Models(LLMs), has introduced a new paradigm: content creation. In cybersecurity, this capability is being harnessed to create a new generation of intelligent tools that can synthesise information, generate code, and simulate threats. GenAI is not merely an incrementalimprovement; it is fundamentally changing the nature of human-computer interactionwithin security operations, acting as a co-pilot and a force multiplier for defenders.

4.1. Benefits

GenAI adds language-native reasoning and generation that compresses analyst cycle time for triage, research, and communication

- **Accelerated Analysis and Synthesis**: LLMs ingest unstructured threat intel, reports, forums, and logs, producing concise, actionable summaries in natural language.
- **Democratisation of Security Expertise**: Co-pilots translate intent to queries, detections, or reverse-engineering steps, enabling junior analysts to perform tasks previously limited to specialists.
- **Proactive Defence through Simulation**: Generative approaches produce realistic synthetic attack data for training and red-team validation without exposing sensitive datasets.

4.2. Use Cases

GenAI is being applied across the security lifecycle to augment and automate critical functions.

- **Threat Intelligence and Hunting**: LLMs map CVEs to ATT&CK, draft detections, and extract IOCs/IoBs from unstructured sources with promptable workflows.
- **Automated Incident Response and Reporting**: GenAI accelerates timeline reconstruction, proposes mitigations, and drafts stakeholder-targeted updates that stay consistent under time pressure.
- **Secure Code Development and Remediation**: GenAI suggests secure fixes during coding and explains risks in plain language to embed DevSecOps earlier in the lifecycle.
- **Security Awareness Training**: Personalized, realistic phishing simulations improve resilience by mirroring attacker personalization tactics.

4.2.1. Seqrite Use Cases

Seqrite integrates GenAI into XDR through SIA, a virtual security analyst that provides conversational investigations, predefined prompts, and contextual summaries to accelerate decision-making. GenAI enhancements add automated summaries, contextual threat mapping, natural-language guidance, and workload reduction to mitigate alert fatigue and improve SOC efficiency.

- Natural-language incident investigations with predefined prompts to retrieve incident details, aggregate alerts by severity or MITRE techniques, and identify cross-incident patterns.
- Contextual summarization of incidents with suggested mitigations, trend insights, and deep links to incidents, rules, and playbooks for rapid action.
- Automated incident summaries and analyst assistance that cut workload by up to 50%.
- GenAI-driven alert prioritization and false-positive reduction of 40–70% to reduce fatigue and focus attention on critical issues.
- Contextual threat mapping that correlates alerts with frameworks like MITRE ATT&CK to speed understanding and response.

- Conversational access to documentation, best practices, and step-by-step task guidance to reduce ramp-up time and streamline reporting.
- Real-time threat hunting support using IOCs and MITRE TTP-based rules within a unified XDR workbench for faster pivots and deeper investigations.

[9], [10], [11]

4.3. Case Studies

Leading technology and security companies are rapidly integrating GenAI into their platforms to deliver next-generation capabilities.

- **Microsoft Security Copilot**: This tool embeds a GPT-4-based LLM directly into Microsoft's security products (including Sentinel and Defender). It functions as a natural language assistant for security analysts, allowing them to summarise complex incidents, generate threat hunting queries in KQL (Kusto Query Language), analyse malware scripts, and receive guided response recommendations. This empowers analysts to investigate and respond to threats more quickly and efficiently. [12]
- **Google Threat Intelligence**: Google has integrated its powerful Gemini model into its threat intelligence offerings. This allows security professionals to use conversational search to query Google's vast repository of threat data, which includes insights from its Mandiant division and crowdsourced data from VirusTotal. The tool can analyse potentially malicious code and provide a natural language summary of its behaviour, significantly speeding up malware analysis and threat research. [13], [14]
- **Swimlane Turbine (Hero AI)**: Swimlane has integrated GenAI into its SOAR platform. Their "Hero AI" allows analysts to perform tasks like summarising security cases and mapping data schemas using natural language. A key feature is the ability for analysts to create complex automation workflows by simply describing the desired process in English, which the AI then translates into an executable playbook. One early adopterreported that the tool enabled their team to close 5,000 cases in a remarkably short period. [15]

5. Implications of AI/ML on Cybersecurity

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into cybersecurity represents more than a technological upgrade; it is a fundamental disruption that carries profound implications for both defenders and attackers. AI is a dual-use technology, and its capabilities are being weaponised as effectively as they are being used for defence. This has created a new set of challenges, empowered adversaries with novel attack vectors, and fundamentally altered the strategic balance of the cyber domain.

5.1. New Challenges

As detailed in section 2, GenAI has scaled phishing, social engineering, and polymorphic malware while lowering the skill barrier, which increases attack volume and pace against defenders.

- **Increased Volume and Velocity of Attacks**: AI enables adversaries to automate and scale their operations to an unprecedented degree. Tasks that once required significant manual effort, such as reconnaissance or crafting phishing emails, can now be executed by AI at machine speed. This results in a dramatic increase in the volume and velocity of attacks, threatening to overwhelm conventional security defences and human-led Security Operations Centres (SOCs).
- **Hyper-Personalised and Evasive Threats**: Generative AI allows for the creation of highly sophisticated and evasive threats. Attackers can now generate phishing emails that are not only grammatically perfect but also highly personalised, incorporating details scraped from a target's social media profiles to make them exceptionally convincing. Similarly, AI can be used to create polymorphic malware that continuously alters its code to evade signature-based detection systems.
- **Lowered Barrier to Entry for Attackers**: Perhaps one of the most significant implications is the democratisation of advanced cybercrime. AI tools, particularly Large Language Models (LLMs), have lowered the technical skill required to launch sophisticated attacks. Novice actors can now use AI to write malicious code, identify vulnerabilities, and orchestrate complex campaigns that previously would have been the exclusive domain of highly skilled, state-sponsored groups. This expansion of the threat actor landscape increases the overall risk for all organisations.

5.2. Attacker Use Cases

Adversaries are actively integrating AI and ML into every stage of the attack lifecycle to enhance their efficiency and effectiveness

- **AI-Powered Reconnaissance**: Attackers use AI to automate the initial phase of an attack. AI-driven tools can scan vast networks for vulnerabilities, identify misconfigurations, and gather intelligence on potential targets from public sources like corporate websites and social media. This allows for the rapid and automated creation of detailed target profiles to be used in subsequent attack stages.
- **Generative AI for Social Engineering**: This is one of the most immediate and impactful malicious uses of AI. LLMs are used to craft highly convincing phishing emails, text messages, and social media posts at scale. Beyond text, deepfake technology allows attackers to create realistic audio and video clones of trusted individuals, such as a CEO or a family member, to manipulate victims into authorising fraudulent transactions or divulging sensitive information.
- **AI-Generated Malware and Exploits**: While AI is not yet capable of autonomously creating entirely novel zero-day exploits, it serves as a powerful assistant for malware developers. GenAI can generate code snippets for specific malicious functions, create variations of existing malware to bypass antivirus detection (polymorphism), and help attackers find and weaponise known vulnerabilities more quickly. Research from MIT Sloan and Safe Security found that 80% of modern ransomware attacks now incorporate AI in some form. [16]
- **Adversarial Attacks on Defensive AI**: A sophisticated new vector involves attacking the defensive AI systems themselves. Adversaries use techniques like data poisoning to corrupt the training data of an ML model, creating backdoors or blind spots. They also use evasion attacks, where a malicious input is subtly modified to be misclassified as benign by an AI detector. This represents a shift from attacking the network to attacking the security logic that protects it.

5.3. Examples

- **WormGPT and FraudGPT**: In 2023, malicious LLMs began appearing on dark web forums. Tools like WormGPT and FraudGPT are versions of open-source models that have been trained without the ethical safeguards present in commercial offerings. They are explicitly marketed to cybercriminals as tools for generating malicious code, writing convincing phishing emails, and creating content for fraud schemes. [17]

Deepfake Voice Scams: There have been multiple documented cases of attackers using AI-generated voice clones to conduct fraud. In one high-profile incident, fraudsters used a deepfake of a CEO's voice to convince a senior executive to transfer hundreds of thousands of dollars to a fraudulent account. The voice was so convincing that the executive believed they were following a direct order from their superior. [18]

Automated Vulnerability Exploitation: The potential for AI to accelerate the exploitation of vulnerabilities has been demonstrated in academic research. A 2024 study showed that GPT-4, when provided only with the public CVE description of a vulnerability, was able to successfully exploit 87% of the critical-severity vulnerabilities tested. This suggests that AI could dramatically shorten the window between the disclosure of a vulnerability and its widespread exploitation by attackers. [19]

The introduction of AI is fundamentally reshaping the traditional asymmetry of cybersecurity. Historically, the defender was at a disadvantage, needing to secure every potential point of entry while an attacker needed to find only a single flaw. However, launching a sophisticated, targeted attack required significant time, resources, and expertise, which limited the number of high-level adversaries. Generative AI upends this dynamic. It dramatically lowers the cost and skill required to execute advanced attacks, allowing a single, less-skilled actor to automate reconnaissance, craft thousands of unique, personalised phishing emails, and generate evasive malware variants. This transforms the nature of the asymmetry from a purely tactical one (one vulnerability versus total defence) to a strategic one rooted in AI capability. The advantage now shifts to the side that can develop, deploy, and adapt its AI models more effectively and at a greater scale, turning cybersecurity into a direct confrontation between the attacker's AI and the defender's AI.

6. Challenges in AI/ML due to Lack of Cybersecurity

While the previous chapter detailed howadversaries leverage AI to attack organisations, this chapter focuses on the inverse: the vulnerabilities and challenges that arise within AI and ML systems themselves when they are not properly secured. As AI models become integral to critical business and security functions, they have become a high-value target. A lack of cybersecurity for the AI pipeline itself can lead to model manipulation, data privacy breaches, and catastrophic failures in detection, creating a new and complex set of risks that organisations must manage.

6.1. MITRE ATLAS Framework

To provide a structured way to understand threats against AI systems, MITRE developed the Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) framework. Analogous to the well-known ATT&CK framework for traditional cybersecurity threats, ATLAS is a knowledge base of adversary tactics and techniques based on real-world observations and red-teaming exercises targeting the ML lifecycle.

The ATLAS framework is organised into tactics that represent the adversary's high-level objectives. Key tactics include:

- **Reconnaissance**: The adversary gathers information about the target AI model, its architecture, and its training data to identify potential weaknesses.
- **ML Model Access**: The adversary gains access to a proprietary model, either through theft or by exploiting an insecure API, to enable further attacks like model extraction or evasion.
- **Evasion**: The adversary crafts specific inputs at inference time that are designed to be misclassified by the model. A classic example in cybersecurity is slightly modifying a malware sample so that an AI-based antivirus system classifies it as benign.
- **Poisoning**: The adversary manipulates the training process by injecting malicious data or corrupting existing data. This can be used to create a "backdoor" in the model, causing it to behave in a specific malicious way when it sees a certain trigger, or to simply degrade the model's overall performance.
- **Inference Attacks**: The adversary queries a deployed model to extract sensitive information. This includes membership inference attacks, which aim to determine if a specific individual's data was in the training set, and model extraction attacks, which attempt to reconstruct the model itself.
- **Supply Chain Compromise**: The adversary targets the components used to build the AI system, such as open-source libraries, pre-trained models downloaded from public repositories, or third-party data sources. There are reported instances of GPUs that are used in AI pipeline getting compromised as well.

6.2. Examples of Data Privacy Breaches

The vast quantities of data required to train modern AI models create significant privacy risks, especially when proper security controls are not in place.

- **Training Data Leakage**: LLMs, trained on massive datasets scraped from the internet, can "memorise" sensitive information contained within that data. If not properly sanitised, this information can be revealed in the model's responses. High profile incidents have occurred where corporate employees input proprietary source code or confidential business plans into public LLMs for assistance, inadvertently incorporating that sensitive data into the model's training set for future use. In one notable case, a bug in ChatGPT led to the leakage of user chat histories, exposing conversations to other users.

- **Model Inversion and Inference Attacks**: These attacks represent a more subtle but equally dangerous privacy breach. By carefully crafting a series of queries, an attacker can exploit a model's outputs to infer information about the data it was trained on. For example, an attacker could potentially reconstruct sensitive medical images or personal financial data from a model trained on that information, even without direct access to the original dataset.

6.3. Missing True Positives (False Negatives)

While false positives (incorrectly flagging benign activity as malicious) are a well-known problem that leads to alert fatigue, false negatives (failing to detect a genuine threat) represent a more insidious and dangerous failure of AI-powered security.

- **The Silent Failure**: A false negative is a silent failure. The system does not generate an alert, giving the organisation a false sense of security while a real attack is underway. Over-reliance on AI systems without acknowledging the risk of false negatives can lead to catastrophic breaches, as an attacker who successfully evades detection can operate within a network undetected for an extended period.

- **Causes of False Negatives:**

  Adversarial Evasion: This is a primary cause, where an attacker deliberately crafts a malicious payload to bypass the AI's detection logic.

  Concept Drift: The nature of cyber threats is constantly evolving. If an AI model is not continuously retrained on new data, its knowledge becomes outdated, and it may fail to recognise new attack techniques – a phenomenon known as concept drift.

  Data and Algorithmic Bias: If a model's training data is not diverse and representative of the full spectrum of threats, it will have blind spots. For instance, a model trained primarily on malware targeting Windows systems may perform poorly at detecting threats against Linux or macOS, leading to false negatives for those operating systems.

6.4. Other Use Cases of AI Vulnerabilities

- **The "Black Box" Problem**: Many of the most powerful AI models, especially deep neural networks, operate as "black boxes." Their internal decision-making processes are so complex that they are not easily interpretable by humans. When such a model flags an activity as malicious, it can be difficult for a human analyst to understand whythat conclusion was reached. This lack of explainability hinders incident response, makes it difficult to debug model errors, and erodes trust in the system.
- **Data Quality and Poisoning**: The adage "garbage in, garbage out" is especially true for AI. The performance of any ML model is fundamentally limited by the quality of its training data. If the data is noisy, incomplete, or biased, the resulting model will be unreliable. This dependency on data quality is itself a vulnerability. Attackers can intentionally "poison" a training dataset by injecting carefully crafted malicious data, subtly manipulating the model's behaviour to serve their own ends.

7. Guardrails for Safe and Secure AI/ML

The dual-use nature of Artificial Intelligence (AI) and the emergence of new vulnerabilities targeting AI systems themselves necessitate a structured and disciplined approach to its deployment, guided by community-vetted standards and controls. Simply adopting AI tools for their defensive capabilities is insufficient; organisations must also implement robust governance frameworks and technical safeguards mapped to the OWASP ML Security Top 10 and the OWASP Top 10 for LLM Applications to manage inherent risks from data through deployment.

7.1. OWASP TOP 10

7.1.1. OWASP Machine Learning Security Top Ten

1. I**nput Manipulation Attack**: Attackers craft inputs – often subtle, adversarial perturbations – to induce misclassification or undesired behavior in ML models, enabling evasion of controls like image or intrusion detectors. Mitigations include adversarial training, robust modeling approaches, and strong input validation to detect anomalies before inference.

2. **Data Poisoning**: Threat actors inject or relabel training samples so a model internalizes incorrect patterns or backdoors, degrading accuracy or causing targeted failures post-deployment. Defenses include data validation and verification, access control on datasets, monitoring and auditing of data pipelines, model validation on hold-out sets, and ensembling to reduce single-point compromise.

3. **Model Inversion**: By repeatedly querying a model and analyzing outputs, adversaries can infer sensitive attributes from training data or reconstruct representative inputs. Limiting access to models, validating inputs, logging and monitoring outputs for anomalies, and periodic retraining help reduce leakage and detect abuse.

4. **Membership Inference Attack**: Attackers determine whether a specific record was in the training set, risking sensitive data disclosure and regulatory exposure. Defensive techniques include differential privacy or output obfuscation, regularization to limit overfitting, dataset minimization, and continuous testing and monitoring for abnormal model behavior.

5. **Model Theft / Extraction**: Adversaries replicate or exfiltrate a proprietary model via API queries, reverse engineering, or access to artifacts, causing IP loss and downstream misuse. Controls include encryption at rest and in transit, strict access controls and authentication, code/graph obfuscation, watermarking, backups, and monitoring and auditing model access.

6. **Adversarial Example Attack**: A focused class of input manipulation that introduces small, carefully constructed changes to inputs to force targeted mispredictions while remaining imperceptible to humans. Robustness can be improved with adversarial training, anomaly detection on inputs, and layered validation around the inference boundary.

7. **Model Evasion**: Attackers adapt inputs to bypass ML-based defenses (for example, tuning network traffic so IDS classifiers overlook malicious flows), undermining detection and response. Combining resilient feature sets, input validation, and continuous monitoring for distribution shifts reduces evasion success.

8. **Exploit of ML Software Flaws**: Compromised or vulnerable ML dependencies, packages, or models can introduce malicious behavior into pipelines and runtime environments via the AI supply chain. Verify package signatures, use trusted repositories, keep dependencies updated, apply code reviews, and employ package verification tools to prevent downstream compromise.

- **9. Insecure Model Deployment**: Misconfigured endpoints, weak authentication, or exposed artifacts enable unauthorized access, theft, or misuse of models and data. Enforce strong access controls and authentication, encrypt artifacts, isolate environments, and monitor and audit model usage to detect and prevent abuse.
- **10. Lack of Auditing and Monitoring**: Without telemetry and audit trails across data, training, and inference, organizations cannot detect poisoning, inversion, extraction, or drift in a timely manner. Implement continuous monitoring and auditing recommended across the ML Top 10 (e.g., data pipelines, model outputs, and access patterns) to rapidly surface anomalies and attacks.

[20]

7.1.2. OWASP Top 10 for Large Language Model Applications

1. **Prompt Injection**: Crafted prompts can override intended instructions, leading to unauthorized actions, data exfiltration, or unsafe outputs. This risk targets the model's reasoning layer, enabling policy bypass and compromised decision-making.
2. **Insecure Output Handling**: Treating generated content as trusted can enable downstream exploits such as code execution or injection into connected systems. Validating and sanitizing LLM outputs is essential to prevent propagation of malicious or unsafe content.
3. **Training Data Poisoning**: Tampered corpora seed backdoors, bias, or harmful behaviors that surface at inference, degrading safety and reliability. Compromised training data can manifest as manipulated responses, eroding integrity and trust.
4. **Model Denial of Service**: Adversaries issue resource-intensive prompts or workflows that inflate latency, cost, or cause outages for LLM-backed services. Capacity exhaustion results in degraded availability and operational disruption.
5. **Supply Chain Vulnerabilities**: Untrusted tools, plugins, libraries, models, or datasets compromise LLM application integrity and can cascade into systemic failures. Dependence on compromised components increases breach likelihood and impact.
6. **Sensitive Information Disclosure**: LLMs may surface confidential inputs, training data, or secrets, causing legal, compliance, or competitive harm. Insufficient safeguards around prompts, context, and outputs lead to unintended data exposure.
7. **Insecure Plugin Design**: Plugins processing untrusted inputs with excessive privileges risk severe exploits, including remote code execution. Weak access control and isolation amplify blast radius within agentic or tool-augmented LLM systems.
8. **Excessive Agency**: Granting LLMs autonomous actions without robust guardrails can trigger unintended operations that jeopardize safety and privacy. Over permissioning and weak oversight undermine reliability and trustworthiness.

9. **Overreliance**: Blind trust in generated content leads to poor decisions, security gaps, and liability when outputs are incorrect or fabricated. Human validation and governance are necessary to counter hallucinations and errors.

10. **Model Theft**: Unauthorized access to proprietary LLM weights or parameters results in IP loss and enables adversarial reuse or cloning. Theft undermines competitive advantage and can propagate unsafe or unbounded copies in the wild

21]

7.2. Other Guardrails

7.2.1. AI Trust, Risk, and Security Management (AI TRiSM)

Popularised by the technology research firm Gartner, AI Trust, Risk, and Security Management (AI TRiSM) is a comprehensive framework designed to ensure AI model governance, trustworthiness, fairness, reliability, and data protection. It is not a single product but a holistic program that integrates people, processes, and technology across the entire AI lifecycle.

The AI TRiSM framework is built on several core pillars or functions:

- **Explainability and Model Monitoring**: This pillar addresses the “black box” problem of AI. It requires that organisations can interpret and explain the decisions made by their AI models. This builds trust among users and stakeholders and is crucial for debugging and accountability. It also involves continuous monitoring of models in production to detect performance degradation, data drift, or the emergence of bias over time.
- **ModelOps**: Analogous to DevOps for software, ModelOps provides a streamlined, automated approach to managing the lifecycle of AI models. It covers all stages, from data preparation and model training to deployment, monitoring, and retraining. A mature ModelOps practice ensures that models are deployed efficiently, reliably, and consistently. Further, the ModelOps should be integrated with organisation’s DevSecOps pipeline.
- **AI Application Security**: This function focuses specifically on protecting AI systems from adversarial attacks. It involves implementing defences against techniques like data poisoning, evasion attacks, and model theft, ensuring the integrity and resilience of the AI models themselves.
- **Data Protection and Privacy**: AI models are trained on vast amounts of data, which often includes sensitive or personal information. This pillar mandates robust data governance, ensuring that data is collected, stored, and used securely and in compliance with privacy regulations such as the GDPR. It includes practices like data encryption, access control, and anonymisation.

Implementing AI TRiSM requires a concerted effort, starting with the establishment of a cross-functional governance team that includes representatives from technology, security, legal, and business units. Security and risk management must be integrated into the AI development process from the very beginning, not treated as an afterthought.
[22]

7.2.2. Responsible AI

The concept of Responsible AI extends beyond technical security to encompass broader ethical and societal considerations. Several key frameworks provide principles and guidelines for developing and deploying AI in a responsible manner.

NIST AI Risk Management Framework (AI RMF): Developed by the U.S. National Institute of Standards and Technology (NIST), the AI RMF is a voluntary framework designed to help organisations manage the risks of AI to individuals, organisations, and society. It provides a structured, lifecycle-based approach to responsible AI. The framework is organised around four core functions:

- **Govern**: This is a cross-cutting function that establishes a culture of risk management. It involves creating policies, assigning roles and responsibilities, and ensuring that AI risk management is integrated into the organisation's broader governance structures.
- **Map**: This function focuses on establishing the context and identifying the potential risks associated with an AI system throughout its lifecycle.
- **Measure**: This involves using quantitative and qualitative methods to analyse, assess, and monitor the risks identified in the Map function.
- **Manage**: This function is about treating the identified risks. It involves developing and implementing plans to mitigate, transfer, or accept risks based on the organisation's risk tolerance.

The NIST RMF promotes the development of "trustworthy AI," which it defines by seven key characteristics: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful biases managed.

- **ISO/IEC 42001**: This is the first international standard for an AI Management System (AIMS). It provides a certifiable framework for organisations to manage the quality, security, privacy, and ethics of their AI systems. Its structure is aligned with other ISO management standards (like ISO 27001 for information security), making it easier to integrate into existing governance processes. Key principles of ISO 42001 include a risk-based approach, lifecycle management, ethical considerations such as fairness and transparency, and robust data governance to ensure data protection.

A common thread runs through these modern governance frameworks: they all advocate for a continuous, lifecycle-based approach to AI security and risk management. This represents a significant departure from traditional security practices, where security controls were often applied as a final step after a system was already built. This shift is a direct response to the unique nature of AI vulnerabilities.

Threats like data poisoning, algorithmic bias, and privacy leaks are not vulnerabilities that can be patched after deployment; they are often embedded deep within the training data and model architecture during the development process itself. Therefore, frameworks like AI TRiSM, the NIST AI RMF, and ISO 42001 all mandate that governance, risk assessment, and security

controls be integrated into every stage of the AI lifecycle – from initial data sourcing and model design to deployment, monitoring, and eventual retirement. This "shift-left" approach is no longer a best practice but a fundamental necessity for building genuinely secure and trustworthy AI systems.

7.3. Free Tools for Testing

Mentioned below are some reputable free, open-source tools to test AI/ML and LLM systems across adversarial robustness, red teaming, evaluation, and guardrail enforcement.

- **IBM Adversarial Robustness Toolbox (ART)**: A Python library to evaluate, defend, and verify ML models against adversarial threats, supporting evasion, poisoning, extraction, and inference attacks across major ML frameworks and data types. ART's documentation highlights extensive attack and defense modules plus metrics for robustness, certification, and verification to systematically test models. [23]
- **Microsoft Counterfit**: A command-line, model-agnostic automation layer for red teaming AI systems that enables pen testing, vulnerability scanning, and attack logging against models across environments, based on Microsoft's internal AI red team practices. Official descriptions emphasize simulating published adversarial attacks at scale to assess robustness before production. [24]
- **garak (LLM vulnerability scanner)**: An open-source CLI that probes LLMs for failure modes such as hallucination, data leakage, prompt injection, misinformation, and toxicity to support generative AI red teaming and assessment. The official site and repository detail install options and active probes to systematically uncover LLM weaknesses. [25]
- **Promptfoo**: An open-source framework to test prompts, agents, and RAG pipelines with scoring, red teaming, and CI/CD workflows, including a maintained GitHub Action for automated evaluations in pull requests. The official docs and action repo provide configuration patterns to run prompt evaluations on every change for continuous quality and safety checks. [26]
- **NVIDIA NeMo Guardrails**: A toolkit to define and orchestrate guardrails – content safety, topic control, PII detection, RAG enforcement, and jailbreak prevention – for safer agentic and conversational LLM applications, with examples and evaluation resources in official docs. Developer materials explain Colang-based rule design and integration with popular LLM frameworks to enforce runtime constraints and security policies. [27]
- **Guardrails AI**: A Python library to specify structure, types, and validators for LLM outputs via the RAIL spec, adding automatic validation and corrective actions (e.g., re asking) to improve safety and reliability of generated content. The official repository and docs describe schema enforcement and semantic validators that help prevent unsafe or malformed outputs propagating downstream. [28]

- **TextAttack**: A Python framework for generating adversarial examples, data augmentation, and adversarial training in NLP, enabling systematic attacks and robustness testing for text models via CLI or Python APIs. Official docs and the EMNLP demo paper present modular attack components and workflows to benchmark and harden NLP systems. [29]
- **Foolbox**: A Python toolbox for creating adversarial examples against models in PyTorch, TensorFlow, and JAX, used to stress-test robustness with a variety of attack methods and reproducible setups. The official site and docs outline installation and usage to integrate adversarial evaluations into model test suites. [30]
- **CleverHans**: A long-standing adversarial example library for constructing attacks, building defenses, and benchmarking vulnerability, with documentation for setup and supported configurations. Official docs describe benchmarking guidance and integration patterns to evaluate susceptibility and defense effectiveness. [31]

8. Approach to Implement Safe and Secure AI/ML in Cybersecurity

Implementing safe and secure Artificial Intelligence (AI) and Machine Learning (ML) in cybersecurity is not merely a technical challenge; it is a strategic endeavour that requires a structured approach. Organisations must move beyond ad-hoc adoption of AI tools and instead develop a mature, systematic program for managing AI capabilities and risks. This involves assessing their current maturity level, establishing a clear roadmap for advancement, and defining a set of robust metrics to measure performance, effectiveness, and return on investment.

8.1. AI/ML Maturity Models

AI/ML maturity models provide a framework for organisations to benchmark their current capabilities and chart a course for improvement. These models typically define a series of stages, each representing a greater degree of integration, automation, and strategic value from AI. They help organisations take the guesswork out of security planning by providing a structured way to identify gaps, set priorities, and track progress over time.

A prominent example is the **AI Maturity Model for Cybersecurity** developed by Darktrace. This model evaluates an organisation's maturity across five levels, considering three key dimensions at each level: the cybersecurity outcomes achieved, the evolving role of human personnel, and the sophistication of the underlying technology

00

• **Level 0 – Manual Operations**: Security processes are almost entirely manual and reactive. Analysts manually investigate alerts from basic tools, and most alerts go uninvestigated due to high volume. Threat hunting is sporadic and relies on human effort.

01

• **Level 1 – Automation Rules**: Organisations use basic automation, such as Security Orchestration, Automation, and Response (SOAR) playbooks or scripts, to handleknown threats and repetitive tasks. However, this automation is rigid, requires constant manual tuning, and is ineffective against novel attacks.

02

• **Level 2 – AI Assistance**: AI tools, particularly Generative AI co-pilots, are introduced to assist analysts with tasks like research, alert summarisation, and query generation. This reduces some manual workload but requires heavy human oversight due to the potential for AI errors and "hallucinations." Trust in autonomous decision-making is low.

03

• **Level 3 – AI Collaboration**: At this level, a true partnership between human and machine emerges. Specialised, purpose-built AI systems are trusted to perform full investigations of alerts and recommend specific response actions. Human analysts transition to a role of oversight, validating high-risk decisions and focusing on strategic tasks like improving detection models and AI governance.

04

• **Level 4 – AI Delegation**: This is the most mature stage, where specialised AI agent systems operate with a high degree of autonomy. The AI handles the majority of security tasks independently at machine speed, including threat detection, investigation, and containment. The human team's role evolves to one of high-level strategic oversight, managing the AI systems, setting policies, and handling only the most critical and exceptional incidents.

By assessing their position on this spectrum, organisations can create a realistic roadmap for AI adoption, focusing on incremental improvements that deliver measurable value at each stage

8.2. Measures & Metrics

To justify investments and manage performance, the impact of AI in cybersecurity must be quantified. Traditional cybersecurity metrics are often insufficient, as they fail to capture the proactive and efficiency-enhancing benefits of AI. A modern, AI-centric approach requires a new set of Key Performance Indicators (KPIs).

Key metrics for evaluating AI in a Security Operations Centre (SOC) include:

- **Mean Time to Detect (MTTD)**: This measures the average time elapsed between the start of a security incident and its detection. AI should significantly reduce MTTD because its algorithms can analyse data and identify anomalies far faster than human analysts. Tracking this metric before and after AI implementation provides a clear measure of its impact on detection speed.
- **Mean Time to Respond (MTTR)**: This metric tracks the average time taken to contain and remediate an incident after it has been detected. AI-powered automation, such as triggering response playbooks in a SOAR platform, can drastically reduce MTTR by executing containment actions in seconds.
- **False Positive Rate (FPR)**: False positives are a major drain on SOC resources, with analysts spending up to a third of their time investigating non-threatening alerts. A key goal of a well-tuned AI system is to reduce the FPR by learning to better distinguish between genuine threats and benign anomalies. A lower FPR is a direct indicator of improved AI accuracy and efficiency.
- **Alert Fatigue Reduction**: While related to FPR, this metric also captures the qualitative impact on the SOC team. It can be measured through surveys assessing analyst satisfaction and perceived workload, as well as by tracking the overall volume of alerts requiring manual investigation. A successful AI implementation should lead to a noticeable reduction in analyst burnout.
- **Model Performance Metrics**: For specific ML models (e.g., a malware classifier), standard performance metrics must be tracked rigorously during testing and continuous monitoring. These include **Accuracy** (overall correct predictions), **Precision** (proportion of positive identifications that were actually correct), **Recall**(proportion of actual positives that were correctly identified), and the **F1-Score** (a harmonic mean of precision and recall).
- **Return on Investment (ROI)**: Extensive use of security AI and automation correlates with materially lower breach costs; align analysis to the earlier USD 1.9 million figure to keep the paper internally consistent.

[32]

9. Conclusion

The integration of ArtificialIntelligence and Machine Learning intocybersecurity marks an irreversible and transformative shift. As this whitepaper has detailed, AI is not merely another tool in the defender's arsenal; it is a foundational technology that is reshaping the principles of digital conflict, creating unprecedented opportunities for defence while simultaneously arming adversaries with formidable new capabilities. The journey from static, rule-based systems to dynamic, self-learning, and now generative AI defences illustrates a rapidly accelerating arms race where the pace of adaptation is paramount.

The key findings of this analysis converge on a central theme: the dual-use nature of AI necessitates a dual-pronged strategy. Organisations can no longer focus solely on leveraging **AI for security** – using its power for threat detection, intelligence analysis, and automated response. They must now equally prioritise the **security of AI** – protecting their own AI models and data pipelines from a new class of threats, including data poisoning, evasion attacks, and sensitive information leakage, as systematically outlined by frameworks like MITRE ATLAS.

The rise of Generative AI has amplified this duality. While it empowers security teams to analyse threats and automate responses with remarkable speed and efficiency, it also provides adversaries with the means to craft highly sophisticated and scalable attacks, effectively democratising cybercrime. This has collapsed the innovation-exploitation cycle, transforming the strategic landscape from a static defence of perimeters to a dynamic, real-time contest of competing AI systems

Navigating this new reality requires a deliberate and mature approach to AI adoption. The path forward for any organisation serious about cybersecurity must be guided by several core principles:

1. **Adopt a Lifecycle Approach to AI Governance**: Security and risk management cannot be an afterthought. Frameworks like Gartner's AI TRiSM, the NIST AI RMF, and ISO/IEC 42001 all underscore the necessity of embedding security, ethics, and trust into every stage of the AI lifecycle, from data sourcing and model development to deployment and continuous monitoring.
2. **Augment, Do Not Replace, Human Expertise**: The ultimate goal of AI in cybersecurity should be to augment the capabilities of human analysts, not to replace them. AI should handle the scale and speed of data analysis, freeing human experts from repetitive tasks to focus on strategic oversight, complex threat hunting, and the governance of the AI systems themselves. The human-in-the-loop remains the most resilient model.
3. **Measure What Matters**: The effectiveness of AI must be quantified through meaningful metrics. Organisations should move beyond traditional measures and track KPIs that reflect AI's true impact, such as reductions in Mean Time to Detect (MTTD), Mean Time to Respond (MTTR), and false positive rates, alongside a clear calculation of return on investment.
4. **Foster Continuous Learning and Adaptation**: The AI-driven threat landscape is not static; it is a fluid, rapidly evolving environment. An organisation's security posture must be equally dynamic. This requires a commitment to continuous learning, regular model retraining, and the use of maturity models to guide a perpetual cycle of assessment and improvement.
5. **Supply Chain Assessment**: Inventorise all hardware and software AI assets including open source for vulnerability scanning, DevSecOps and patch management.
6. **Zero Trust Architecture**: Organisations should implement security controls in a wholistic approach and stick to the fundamentals of Zero Trust like any other IT assets including SOC monitoring and sharing threat intelligence of AI assets.

In conclusion, the future of cybersecurity is inextricably linked with the future of AI. Success will not be determined by the mere possession of AI tools, but by the wisdom, discipline, and strategic foresight with which they are implemented. By embracing a holistic, risk-informed, and lifecycle-aware strategy, organisations can harness the transformative power of AI to build a more resilient and secure digital future.

10. References

[1] “Intrusion-Detection Research at SRI’s Computer Science Laboratory.” [Online]. Available: https://www.csl.sri.com/programs/intrusion/history.html

[2] “The History of Digital Spam.” [Online]. Available: https://cacm.acm.org/research/the-history-of-digital-spam/

[3] “Stanislav Petrov (U.S. National Park Service).” [Online]. Available: https://www.nps.gov/people/stanislav_petrov.htm

[4] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath, “Malware images: visualization and automatic classification,” in Proceedings of the 8th International Symposium on Visualization for Cyber Security, in VizSec '11. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2011. doi: 10.1145/2016904.2016908.

[5] “Cost of a data breach 2025 | IBM.” [Online]. Available: https://www.ibm.com/reports/data-breach

[6] Seqrite, “Extended Detection and Response | XDR vendors for cybersecurity.” [Online]. Available: https://www.seqrite.com/extended-detection and-response-xdr/

[7] Seqrite, “Why AI Assistance in SecOps is Your Missing Security Shield.” [Online]. Available: https://www.seqrite.com/blog/ai-in-secops-security-shield/

[8] “CordenPharma | DarkTrace Cybersecurity Case Study.” [Online]. Available: https://www.darktrace.com/customers/cordenpharma

[9] J. Karlekar, “5 Benefits of Generative AI in Extended Detection and Response (XDR).” [Online]. Available: https://www.seqrite.com/blog/generative-aixdr-benefits-cybersecurity/

[10] J. Karlekar, “Revolutionizing XDR with Gen AI: Next-Level Security Analysis for Advanced Threat Protection.” [Online]. Available: https://www.seqrite.com/blog/revolutionizing-xdr-gen-ai-cybersecurity-seqrite/

[11] Seqrite, “Seqrite Intelligent Assistant (SIA) - Virtual Security Analyst.” [Online]. Available: https://www.seqrite.com/sia/

[12] Mjcaparas, “Microsoft Security Copilot documentation.” [Online]. Available: https://learn.microsoft.com/en-us/copilot/security/

[13] “Google Threat Intelligence.” [Online]. Available: https://cloud.google.com/security/products/threat-intelligence

[14] “Google Threat Intelligence.” [Online]. Available: https://gtidocs.virustotal.com/

[15] L.-C. S. A. & SOAR Platform and Swimlane, “Swimlane Hero AI Actions Power AI Incident Response.” [Online]. Available: https://swimlane.com/news/swimlane hero-ai-actions/

[16] “80% of ransomware attacks now use artificial intelligence | MIT Sloan.” [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/80-ransomware-attacks now-use-artificial-intelligenc

[17] B. Toulas, “Cybercriminals train AI chatbots for phishing, malware attacks,” BleepingComputer, Jan. 2024, [Online]. Available:https://www.bleepingcomputer.com/news/security/cybercriminals-train-ai-chatb ots-for-phishing-malware attacks/

[18] J. Damiani, “A voice deepfake was used to scam a CEO out of $243,000.” [Online]. Available: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-s cam-a-ceo-out-of-243000/

[19] R. Fang, R. Bindu, A. Gupta, and D. Kang, “LLM Agents can Autonomously Exploit One-day Vulnerabilities.” [Online]. Available: https://arxiv.org/abs/2404.08144

[20] “OWASP Machine Learning Security Top 10 - Draft release v0.3.” [Online]. Available: https://mltop10.info/

[21] O. T. 10 for LLM & Generative AI Security, OWASP Top 10 for LLM & Generative AI Security | LLMRISKS Archive. [Online]. Available: https://genai.owasp.org/llm top-10/

[22] “Tackling Trust, Risk and Security in AI Models.” [Online]. Available: https://www.gartner.com/en/articles/ai-trust-and-ai-risk

[23] “Adversarial Robustness Toolbox (ART).” [Online]. Available: https://github.com/Trusted-AI/adversarial-robustness-toolbox

[24] “Counterfit.” [Online]. Available: https://github.com/Azure/counterfit

[25] “garak: LLM vulnerability scanner.” [Online]. Available: https://garak.ai/

[26] “promptfoo.” [Online]. Available: https://github.com/promptfoo/promptfoo

[27] “NVIDIA NeMo Guardrails.” [Online]. Available: https://docs.nvidia.com/nemo guardrails/index.html

[28] “Guardrails.” [Online]. Available: https://github.com/guardrails-ai/guardrails

[29] “TextAttack Documentation.” [Online]. Available: https://textattack.readthedocs.io/

[30] “Foolbox Documentation.” [Online]. Available: https://foolbox.readthedocs.io/

[31] “CleverHans.” [Online]. Available: https://github.com/cleverhans-lab/cleverhans

[32] “The AI Maturity Model for Cybersecurity | Guided AI Assessment.” [Online]. Available: https://www.darktrace.com/ai-maturity-model

11. Appendix

11.1. Glossary

- **Adversarial Machine Learning**: A field of research and a set of techniques that involve manipulating machine learning models through malicious inputs, with the goal of causing the model to make incorrect predictions or classifications.
- **AI Trust, Risk, and Security Management (AI TRiSM)**: A governance framework, popularised by Gartner, for ensuring that AI models are managed for trustworthiness, fairness, reliability, security, and data protection throughout their lifecycle.
- **Anomaly Detection**: A technique used to identify rare items, events, or observations which raise suspicions by differing significantly from the majority of the data. In cybersecurity, it is used to detect unusual patterns that may indicate a threat.
- **Deep Learning**: A subfield of machine learning based on artificial neural networks with multiple layers (deep neural networks). It is particularly effective at learning patterns from large, unstructured datasets like images and text.
- **Evasion Attack**: A type of adversarial attack where a malicious input is slightly modified at inference time to cause an AI model to misclassify it as benign.
- **Expert System**: An early form of AI system that emulates the decision-making ability of a human expert. It uses a knowledge base of facts and a set of “if-then” rules to solve problems in a specific domain.
- **False Negative**: A test result which incorrectly indicates that a particular condition or attribute is absent. In cybersecurity, it is the failure of a security system to detect a genuine threat.
- **False Positive**: A test result which incorrectly indicates that a particular condition or attribute is present. In cybersecurity, it is an alert that incorrectly flags legitimate activity as malicious, contributing to “alert fatigue.”

- **Generative AI (GenAI)**: A class of AI models, such as Large Language Models (LLMs), that are capable of generating new content, including text, images, code, or synthetic data, based on the patterns learned from their training data.
- **Large Language Model (LLM)**: A type of deep learning model trained on vast amounts of text data, enabling it to understand, generate, and respond to human language in a coherent and contextually relevant manner.
- **MITRE ATLAS**: A globally accessible knowledge base of adversary tactics and techniques against AI-enabled systems, analogous to the MITRE ATT&CK framework for traditional cyber threats.
- **NIST AI Risk Management Framework (RMF)**: A voluntary framework from the U.S. National Institute of Standards and Technology to help organisations manage the risks associated with AI and promote the development of trustworthy and responsible AI systems.
- **Poisoning Attack**: A type of adversarial attack where an attacker manipulates the training data of an AI model to compromise its integrity, for example, by creating a backdoor or degrading its performance.
- **Polymorphic Malware**: Malicious software that can constantly change its identifiable features (like its file name or encryption keys) to evade detection by signature-based security tools.
- **Prompt Injection**: An attack technique against Large Language Models where an attacker crafts malicious input (a “prompt to trick the model into bypassing its safety controls or performing unintended actions.
- **Rule-Based System**: A system that uses a set of predefined rules, typically in an “if then” format, to make decisions. Early firewalls and antivirus software were examples of rule-based systems.
- **Security Orchestration, Automation, and Response (SOAR)**: A category of security tools that allows organisations to collect security data and alerts from various sources, and then use a combination of human and machine power to perform incident response actions.
- **Signature-Based Detection**: A method used by traditional antivirus and intrusion detection systems to identify threats by looking for known patterns or “signatures” of malicious code or activity.
- **Zero-Day Attack**: A cyberattack that occurs on the same day a weakness is discovered in software, before the developer has had time to create a patch or solution to fix it.

11.2. Abbreviations

- **AI**: Artificial Intelligence
- **AIMS**: AI Management System
- **ATLAS**: Adversarial Threat Landscape for Artificial-Intelligence Systems
- **BEC**: Business Email Compromise
- **CNN**: Convolutional Neural Network
- **CVE**: Common Vulnerabilities and Exposures
- **DDoS**: Distributed Denial-of-Service
- **FPR**: False Positive Rate
- **GAN**: Generative Adversarial Network
- **GDPR**: General Data Protection Regulation
- **GenAI**: Generative AI
- **IDES**: Intrusion Detection Expert System
- **IDS**: Intrusion Detection System
- **IOC**: Indicator of Compromise
- **ISO**: International Organization for Standardization
- **KQL**: Kusto Query Language
- **LLM**: Large Language Model
- **ML**: Machine Learning
- **MTTD**: Mean Time to Detect
- **MTTR**: Mean Time to Respond
- **NIDS**: Network Intrusion Detection System
- **NIST**: National Institute of Standards and Technology
- **NLP**: Natural Language Processing

- **PII**: Personally Identifiable Information
- **RMF**: Risk Management Framework
- **RNN**: Recurrent Neural Network
- **ROI**: Return on Investment
- **SIEM**: Security Information and Event Management
- **SOC**: Security Operations Centre
- **SOAR**: Security Orchestration, Automation, and Response
- **SVM**: Support Vector Machine
- **TRiSM**: Trust, Risk, and Security Management
- **TTP**: Tactics, Techniques, and Procedures

11.3. Datasets

Cybersecurity Datasets for AI/ML and GenAI

11.3.1. AI/ML in Cybersecurity Datasets

- **CIC Datasets**: https://www.unb.ca/cic/datasets/index.html
- **IMPACT (DHS Information Marketplace)**: https://www.dhs.gov/science-and technology/impact
- **National Vulnerability Database (NVD)**: https://nvd.nist.gov/
- **MITRE ATT&CK®**: https://attack.mitre.org/
- **IEEE Dataport**: https://ieee-dataport.org/

11.3.2. GenAI in Cybersecurity Datasets

- **CyberMetric Benchmark**: https://arxiv.org/abs/2310.16568
- **Hugging Face Datasets (Cybersecurity Search)**: https://huggingface.co/datasets?search=cybersecurity
- **BCCC Cybersecurity Datasets**: https://www.yorku.ca/research/bccc/ucs-technical/cybersecurity-datasets-cds/

About Seqrite

Seqrite is a leading enterprise cybersecurity solutions provider. With a focus on simplifying cybersecurity, Seqrite delivers comprehensive solutions and services through our patented, AI/ML-powered tech stack to protect businesses against the latest threats by securing devices, applications, networks, cloud, data, and identity. Seqrite is the Enterprise arm of the global cybersecurity brand, Quick Heal Technologies Limited, the only listed cybersecurity products and solutions company in India.

We are the first and only Indian company to have solidified India's position on the global map by collaborating with the Govt. of the USA on its NIST NCCoE's Data Classification project. We are differentiated by our easy-to-deploy, seamless-to-integrate comprehensive solutions providing the highest level of protection against emerging and sophisticated threats powered by state-of-the-art threat intelligence and playbooks backed by world-class service provided by best-in-class security experts at India's largest malware analysis lab – Seqrite Labs. We are the only Indian full-stack company aligned with CSMA architecture recommendations, offering award-winning Endpoint Protection, Enterprise Mobility Management, Zero Trust Network Access, and many more. Seqrite Data Privacy management solution enables organizations to stay fully compliant with the DPDP Act and global regulations.

Today, 30,000+ enterprises in more than 70 countries trust Seqrite with their cybersecurity needs. For more information, please visit: www.seqrite.com

SEQRITE

Solitaire Business Hub, Office No. 7010 C & D, 7th Floor,
Viman Nagar, Pune - 411014, India. www.seqrite.com